MODUL 6

PENGENALAN DATA MINING

6.1. Tujuan

1. Mahasiswa dapat memahami tujuan data mining.
2. Mahasiswa memahami dan mengenal beberapa software aplikasi yang dapat digunakan untuk data mining.

6.2. Landasan Teori

6.2.1. Pengertian Data Mining

Data mining adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basisdata. Informasi yang dihasilkan diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basis data.

Hal-hal yang melatarbelakangi data mining:

1. Melimpahnya data (*overload data*) yang dialami oleh berbagai institusi, perusahaan atau organisasi. Merlimpahnya data ini merupakan akumulasi data transaksi yang terekam bertahun-tahun.
2. Data-data tersebut merupakan data transaksi yang umumnya diproses menggunakan aplikasi komputer yang biasa disebut dengan OLTP (*On Line Transaction Processing*).
3. Adanya ledakan informasi (*explotion information*) dari berbagai media terutama internet yang secara

©Muhammadiyah university Press

umum informasi tersebut tidak terstruktur (*unstructured information*).

6.2.2. Manfaat Penggunaan Data Mining

A. Sudut Pandang Komersial

Data mining dapat digunakan dalam menangani meledaknya *volume* data. Bagaimana menyimpannya, mengestraknya serta memanfaaatkannya. Berbagai teknik komputasi dapat digunakan untuk menghasilkan informasi yang dibutuhkan. Informasi yang dihasilkan menjadi aset untuk meningkatkan daya saing suatu institusi. Data mining tidak hanya digunakan untuk menangani persoalan menumpuknya data / informasi dan bagaimana menyimpannya tanpa kehilangan informasi yang penting (*warehousing*). Data mining juga diperlukan untuk menyelesaikan permasalahan atau menjawab kebutuhan bisnis itu sendiri, sebagai contoh:

1. Bagaimana mengetahui hilangnya pelanggan karena pesaing.
2. Bagaimana mengetahui item produk atau konsumen yang memiliki kesamaan karakteristik.
3. Bagaimana mengidentifikasi produk-produk yang terjual bersamaan dengan produk lain.
4. Bagaimana memprediski tingkat penjualan.
5. Bagaimana menilai tingkat resiko dalam menentukan jumlah produksi suatu item.
6. Bagaimana memprediksi perilaku bisnis di masa yang akan datang.

B. Sudut Pandang Keilmuan

Data mining dapat digunakan untuk meng-*capture*, menganalisis serta menyimpan data yang bersifat *real-time* dan sangat besar, misalnya:

1. Remote sensor yang ditempatkan pada suatu satelit.

©Muhammadiyah university Press

2. Telescope yang digunakan untuk memindai langit.
3. Simulasi saintifik yang membangkitkan data dalam ukuran terabytes.

Data mining merupakan salah satu metode alternatif yang dapat digunakan untuk mengolah data mentah, ketika metode konvensional tidak mungkin untuk dilakukan karena besarnya volume data yang diolah. Hal ini dapat terjadi karena data mining memiliki kemampuan mereduksi data baik melalui teknik katalogisasi, klasifikasi maupun segmentasi.

6.2.3. Proses Data Mining

Data mining sebenarnya merupakan salah satu rangkaian dari proses pencarian pengetahuan dalam database (*Knowledge Discovery in Database* (KDD)). KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interprestasi dan visualisasi dari pola-pola sejumlah kumpulan data. KDD adalah keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola (*pattern*) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti. Serangkaian proses tersebut memiliki tahap sebagai berikut (Tan, 2004):

1. Pembersihan data dan integrasi data (*cleaning and integration*). Proses ini digunakan untuk membuang data yang tidak konsisten dan bersifat noise dari data yang terdapat di berbagai basisdata yang mungkin berbeda format maupun platform yang kemudian diintegrasikan dalam satu *database data warehouse*.
2. Seleksi dan transformasi data (*selection and transformation*). Data yang terdapat dalam *database data warehouse* kemudian direduksi dengan berbagai teknik. Proses reduksi diperlukan untuk mendapatkan hasil yang lebih akurat dan mengurangi waktu komputasi terutama untuk masalah dengan skala besar (*large scale problem*). Beberapa cara seleksi, antara lain:

©Muhammadiyah university Press

a) *Sampling*, yaitu seleksi subset representatif dari populasi data yang besar.

b) *Denoising*, yaitu proses menghilangkan noise dari data yang akan ditransformasikan.

c) *Feature extraction*, yaitu proses membuka spesifikasi data yang signifikan dalam konteks tertentu.

Transformasi data diperlukan sebagai tahap *pre-procecing*, dimana data yang diolah siap untuk ditambang. Beberapa cara transformsi, antara lain (Santosa, 2007):

a) *Centering*, mengurangi setiap data dengan rata-rata dari setiap atribut yang ada.

b) *Normalisation*, membagi setiap data yang di centering dengan standar deviasi dari atribut bersangkutan.

c) *Scaling*, mengubah data sehingga berada dalam skala tertentu.

3. Penambangan data (*data mining*)

   Data-data yang telah diseleksi dan ditransformasi kemudian ditambang dengan berbagai teknik. Proses data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan fungsi-fungsi tertentu. Fungsi atau algoritma dalam data mining sangat bervariasi. Pemilihan fungsi atau algoritma yang tepat sangat bergantung pada tujuan dan proses pencaraian pengetahuan secara keseluruhan.

4. Evaluasi pola dan presentasi pengetahuan

   Tahap ini merupakan bagian dari proses pencarian pengetahuan yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya. Langkah terakhir KDD adalah mempresentasikan pengetahuan dalam bentuk yang mudah dipahami oleh pengguna.

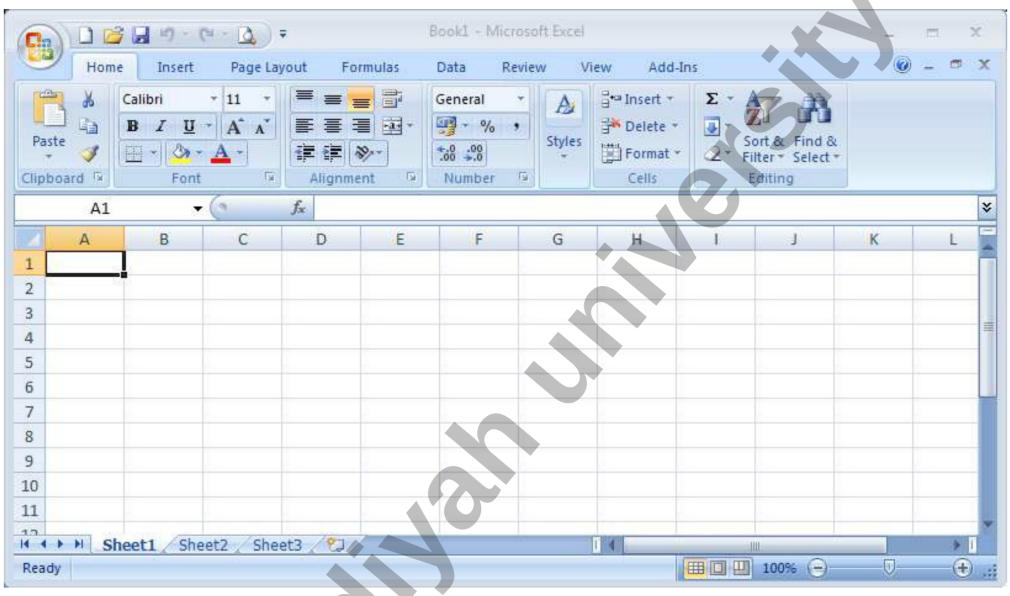
©Muhammadiyah university Press

6.3. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Aplikasi Microsoft Excel, Notepad, Weka, RapidMiner.
3. Modul Praktikum Data Warehousing dan Data Mining.

6.4. Pengenalan Perangkat Lunak Data Mining

6.4.1. Perangkat Lunak 1 : Microsoft Excel



Gambar 6.1 Lembar Kerja Microsoft Excel

Microsoft Excel atau Microsoft Office Excel adalah sebuah program aplikasi lembar kerja spreadsheet yang dibuat dan didistribusikan oleh Microsoft Corporation untuk sistem operasi Microsoft Windows dan Mac OS. Aplikasi ini memiliki fitur kalkulasi dan pembuatan grafik yang, dengan menggunakan strategi marketing Microsoft yang agresif, menjadikan Microsoft Excel sebagai salah satu program komputer yang populer digunakan di dalam komputer mikro hingga saat ini. Bahkan, saat ini program ini merupakan program spreadsheet paling banyak digunakan oleh banyak pihak, baik di platform PC berbasis Windows maupun platform Macintosh berbasis Mac OS, semenjak versi 5.0 diterbitkan

©Muhammadiyah university Press

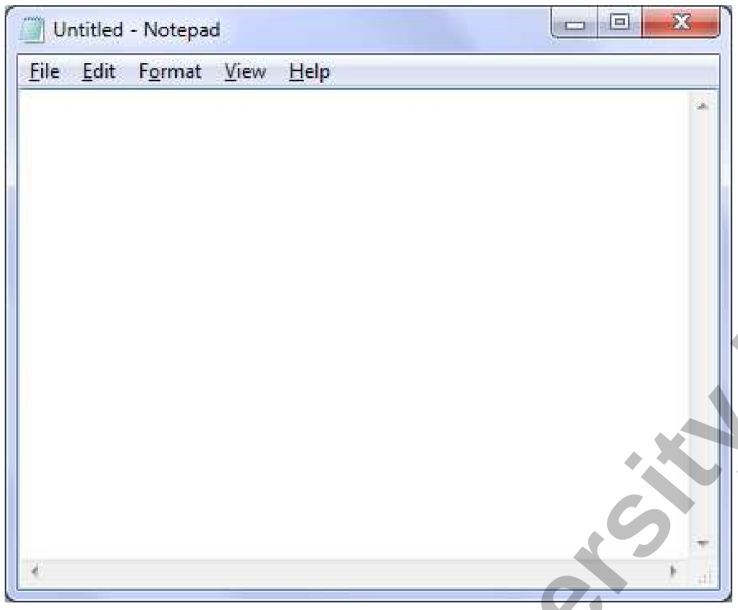
pada tahun 1993. Aplikasi ini merupakan bagian dari Microsoft Office System.

Excel merupakan program spreadsheet pertama yang mengizinkan pengguna untuk mendefinisikan bagaimana tampilan dari spreadsheet yang mereka sunting: font, atribut karakter, dan tampilan setiap sel. Excel juga menawarkan penghitungan kembali terhadap sel-sel secara cerdas, di mana hanya sel yang berkaitan dengan sel tersebut saja yang akan diperbarui nilanya (dimana program-program spreadsheet lainnya akan menghitung ulang keseluruhan data atau menunggu perintah khusus dari pengguna). Selain itu, Excel juga menawarkan fitur pengolahan grafik yang sangat baik.

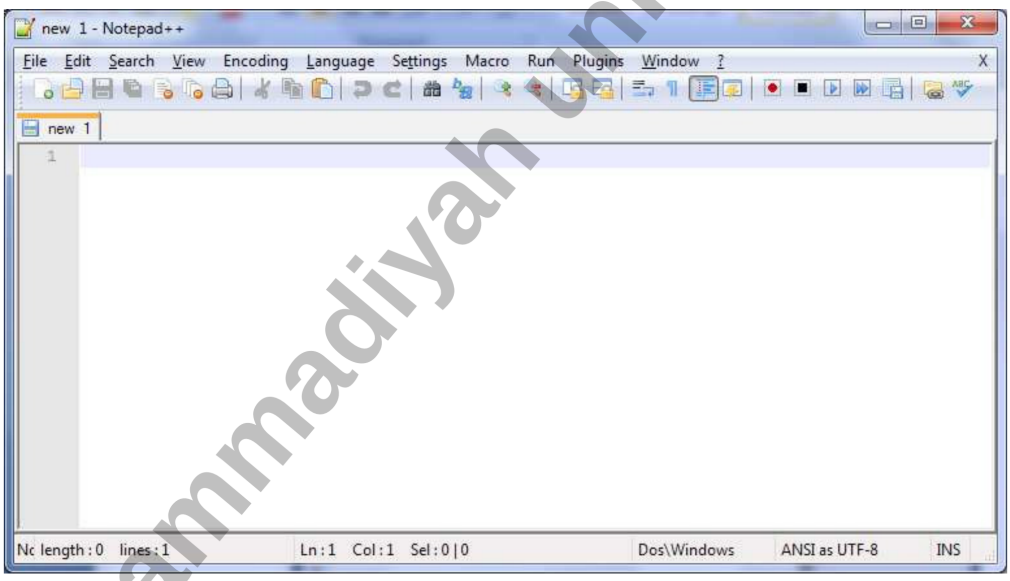
6.4.2. Perangkat Lunak 2 : Notepad / Notepad++

Notepad / Notepad++ adalah sebuah penyunting teks dan penyunting kode sumber yang berjalan di sistem operasi Windows. Notepad adalah salah satu program bawaan dari Windows yang biasa digunakan untuk menulis keterangan-keterangan yang penting dari program aplikasi seperti halnya lisensi program atau yang lainnya. Notepad juga bisa berguna untuk berbagai macam keperluan, seperti membuat file CSS, Javascript untuk format web, dan pembuatan listing pemrograman, seperti Java dan berbagai kegunaan lain. Keuntungan dari penggunaan Notepad adalah kecepatan dan kemudahan dalam pengoperasiannya. Sedangkan dari segi kelemahan, Notepad tidak memiliki tampilan yang menarik. Sementara itu, Notepad++ menggunakan komponen Scintilla untuk dapat menampilkan dan menyuntingan teks dan berkas kode sumber berbagai bahasa pemrograman. Seperti halnya dengan Notepad bawaan Windows, Notepad++ juga bisa digunakan untuk berbagai keperluan pemrograman seperti membuat file CSS, Javascript untuk format web, dan pembuatan listing pemrograman, seperti Java dan berbagai kegunaan lain.

©Muhammadiyah University Press



Gambar 6.2 Lembar Kerja Notepad



Gambar 6.3 Lembar Kerja Notepad++

6.4.3. Perangkat Lunak 3: Weka

WEKA adalah sebuah paket tools machine learning praktis. "WEKA" merupakan singkatan dari Waikato Environment for Knowledge Analysis, yang dibuat di Universitas Waikato, New

©Muhammadiyah University Press

Zealand untuk penelitian, pendidikan dan berbagai aplikasi. WEKA mampu menyelesaikan masalah-masalah data mining di dunia-nyata, khususnya klasifikasi yang mendasari pendekatan-pendekatan machine learning. Perangkat lunak ini ditulis dalam hirarki class Java dengan metode berorientasi objek dan dapat berjalan hampir di semua platform.

Weka terdiri dari koleksi algoritma machine learning yang dapat digunakan untuk melakukan generalisasi / formulasi dari sekumpulan data sampel. Walaupun kekuatan Weka terletak pada algoritma yang makin lengkap dan canggih, kesuksesan data mining tetap terletak pada faktor pengetahuan manusia implementornya. Tugas pengumpulan data yang berkualitas tinggi dan pengetahuan pemodelan dan penggunaan algoritma yang tepat diperlukan untuk menjamin keakuratan formulasi yang diharapkan.

Gambar 6.4 Halaman Depan Weka

Lima tombol pada halaman depan Weka dapat digunakan untuk menjalanankan aplikasi sebagai berikut:

©Muhammadiyah university Press

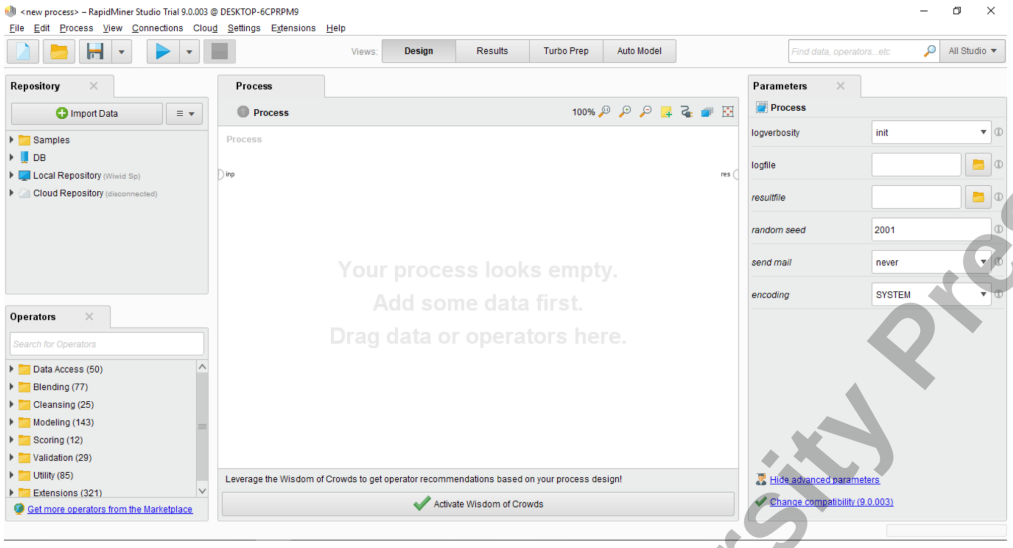
1. **Explorer** digunakan untuk menggali lebih jauh data dengan aplikasi WEKA.
2. **Experimenter** digunakan untuk melakukan percobaan dengan pengujian statistik skema belajar.
3. **KnowledgeFlow** digunakan untuk pengetahuan pendukung.
4. **Workbench**
5. **Simple CLI** antar muka dengan menggunakan tampilan command-line yang memungkinkan langsung mengeksekusi perintah weka untuk Sistem Operasi yang tidak menyediakan secara langsung.

6.4.4. Perangkat Lunak 4 : RapidMiner

RapidMiner adalah sebuah platform perangkat lunak yang dikembangkan oleh perusahaan yang menyediakan lingkungan secara terintegrasi untuk machine learning, data mining, text mining, analisis prediktif dan analisis bisnis. Aplikasi ini digunakan untuk kepentingan bisnis dan industri serta untuk penelitian, pendidikan, pelatihan, prototyping secara cepat, dan pengembangan aplikasi dan mendukung proses data mining termasuk hasil visualisasi, validasi dan optimasi. RapidMiner dikembangkan dari versi sebelumnya yang tersedia di bawah lisensi open source OSI-certificated.

RapidMiner menyediakan prosedur data mining dan machine learning termasuk untuk proses ETL (extraction, transformation, loading), data preprocessing, visualisasi, modeling dan evaluasi. Proses data mining tersusun atas operator-operator yang nestable, yang dideskripsikan dengan XML dan dibuat dengan GUI. Aplikasi ini ditulis dalam bahasa pemrograman Java dan mengintegrasikan proyek data mining Weka dan statistika R.

©Muhammadiyah University Press



Gambar 6.5 Perspektif RapidMiner

©Muhammadiyah university Press

6.5. Tugas (Dikerjakan Saat Ini)

1. Dengan menggunakan Ms. Excel, buatlah tabel berikut.

Jurusan_SMA	Gender	Asal_Sekolah	Rerata_SKS	Asisten	Lama_Studi
IPS	WANITA	SURAKARTA	18	TIDAK	TERLAMBAT
IPA	PRIA	SURAKARTA	19	YA	TEPAT
LAIN	PRIA	SURAKARTA	19	TIDAK	TERLAMBAT
IPA	PRIA	LUAR	17	TIDAK	TERLAMBAT
IPA	WANITA	SURAKARTA	17	TIDAK	TEPAT
IPA	WANITA	LUAR	18	YA	TEPAT
IPA	PRIA	SURAKARTA	18	TIDAK	TERLAMBAT
IPA	PRIA	SURAKARTA	19	TIDAK	TEPAT
IPS	PRIA	LUAR	18	TIDAK	TERLAMBAT
LAIN	WANITA	SURAKARTA	18	TIDAK	TEPAT
IPA	WANITA	SURAKARTA	19	TIDAK	TEPAT
IPS	PRIA	SURAKARTA	20	TIDAK	TEPAT
IPS	PRIA	SURAKARTA	19	TIDAK	TEPAT
IPA	PRIA	SURAKARTA	19	TIDAK	TEPAT
IPA	PRIA	LUAR	22	YA	TEPAT
LAIN	PRIA	SURAKARTA	16	TIDAK	TERLAMBAT
IPS	PRIA	LUAR	20	TIDAK	TEPAT
LAIN	PRIA	LUAR	23	YA	TEPAT
IPA	PRIA	SURAKARTA	21	YA	TEPAT
IPS	PRIA	SURAKARTA	19	TIDAK	TERLAMBAT

2. Dengan menggunakan formula dalam Ms. Excel, carilah:
   a) Pada atribut Jurusan_SMA, berapa jumlah data masing-masing kelas IPA, IPS dan LAIN? (gunakan formula = COUNTIF)
   b) Pada atribut Lama_Studi, berapa jumlah data masing-masing kelas TEPAT, TERLAMBAT?
   c) Pada atribut Rerata_SKS, berapa nilai Max, Min, Mean, dan Standard Deviation?
   d) Pada tabel tersebut, berapakah jumlah data gabungan untuk kelas pada atribut Jurusan_SMA = IPA, Gender = PRIA, Asisten = YA, Lama_studi = TEPAT? (gunakan formula = COUNTIFS)

©Muhammadiyah University Press

3. Dikerjakan di rumah, install 4 buah perangkat lunak dalam modul ini pada komputer anda!

©Muhammadiyah university Press